Multimodal Turn-Taking: Motivations, Methodological Challenges, and Novel Approaches

Katharina J. Rohlfing, Giuseppe Leonardi, Iris Nomikou,
Joanna Rączaszek-Leonardi, and Eyke Hüllermeier (*Senior Member, IEEE)*

***Abstract*—In this article, we note that despite being a multimodal phenomenon, turn-taking has still been investigated mostly as being unimodal. Based on theoretical positions emphasizing that communication is organized jointly by interaction partners, we identify the challenge of assessing human sequential behavior that is (a) spread across different modalities and (b) co-constructed with a partner. By analyzing a corpus of mother–child dyads with Cross Recurrence Quantification Analysis and Frequent Pattern Mining, we offer novel steps toward understanding multimodal turn-taking.**

***Index Terms*—communication, turn-taking, contingency, interaction formats**

I. Introduction

When we consider the phenomenon of *turn-taking*, we think of conversations among adults: When one person is speaking, the other listens, and the other will take her turn when the first speaker finishes her utterance and/or signals a change of a turn by, for example, raising a question. Turn-taking behavior is a universal phenomenon to be found in all cultures; it can be characterized by a rapid exchange of short units of talking [1], [2]. The basic properties of this universal turn-taking system are that (a) individual contributions to the conversational communication – the turns – are units "of no fixed size, but tend to be short, about 2 s in length on average"; and (b) it organizes exchange between the partners so that the overlap between contributions is minimal [2]. This minimal overlap seems to vary across cultures [3], and, moreover, some overlap can also be observed in interaction with infants [4]. Fundamental insights into the systematic organization of communicational exchange stem from microanalytic research in conversational analysis [5] for which the turn is a basic unit of investigation.

Interestingly, the organization of an exchange in turns is not restricted to communication alone. As pointed out by Schegloff [6], this form of exchange extends to any social interaction such as games, traffic at an intersection, and so forth. In his review, Levinson [2] uses the term *duetting* to refer to animal behavior that reveals a pattern similar to the human turn-taking system. Henry and colleagues [7] report such coordination in animal vocal interactions with, for example, starlings favoring alternation over overlap. This similarity with other species contributes to the argument that the turn-taking phenomena in dyadic systems are biological in nature. Stevanovic and Peräkylä [8] consider two functions of such systems: emotional reciprocity and experience sharing. From a developmental perspective, these two functions allow partners to create emotional attunement as a basis for both joint experiences and the transfer of knowledge [9]. Beyond the values that the system yields for learning, one can argue that a more basic function is to coordinate the complementary co-actions of the partners to a dialogue in a form of a sequence. Gratier and colleagues [10], also [11] demonstrate that early vocal exchange between infant and mother has a turn-taking format. In the literature, these early vocal alternations have been dubbed *proto-conversations* [12], [13] and are found to relate closely to children's later phonological development [11].

Nonetheless, a sequence of social actions involves far more than just vocal behavior. In adults, research has shown that gaze serves the important function of nominating the next speaker [14]. Goodwin [15] also showed how interlocutors use gaze in turn construction. In his detailed analysis of pauses, hesitations, and breakdowns in conversations, he showed how participants use gaze to coordinate interaction in dialogue. Building upon this observation, Rutter and Durkin [16] showed a developmental increase from 18 months of age onward in children's use of looks at the end of turns. Methodologically, this study remains an exception because it considered the multimodality of conversational mechanisms and investigated the use of nonverbal behaviors during vocal turns in children. This gap in research is surprising, because the use of nonverbal modalities is at the core of communication and language development. In fact, one of the earlier studies on turn-taking in 2-week-old infants performed by Kaye [17] considered mother–child actions during breastfeeding. He analyzed why infants interrupt their bursts by pauses that do not seem to have any obvious physiological function. The analysis revealed that when interrupting a burst, infants elicit a reaction from their mother in the form of jiggling. One might argue that the jiggling movement would cause the infants to suck again. However, in contrast, the "cessation of jiggling proved to be a better elicitor of a resumption of sucking than the jiggling itself", strongly suggesting that this phenomenon is about exchanging turns rather than "mothers' anticipation of the burst" [18, p. 29]. Interestingly, a change of communicative means in this exchange was noticed with older infants who are by then able to vocalize in the form of coos during the pauses. The maternal reaction is then inserted into the gaps in the baby's activity [11]. The observation is that during an infant's development, the "mother's behavior barely changes: What changes is that the

turn-taking becomes more symmetrical, the baby's turns become real speech acts" [19, p. 212].

Clearly, the basis for this communicative exchange seems to be the infant's sensitivity to regularities and contingencies. The term *contingency* refers to "the conditional probability structure of the contingent relations between responses and stimulus events" [20, p. 102]. It is commonly used as a synonym for turn-taking [21]. Yet, contingent relations are not only of a temporal and sensory kind; they are also spatial. Nagai [22] considers contingency detection to be essential in cognitive development, more specifically, in the development of self- and other recognition. It is a sensitivity that seems to become established between the first and third months of life [23]. Whereas in their first month, infants barely perceive a difference between differently contingent interactions, by the age of 3 months, they can discriminate between them. Gergely and Watson [20, p. 117] postulate that a "contingency detection module" has then become active.

The contingent behavior occurs on different occasions and within various modalities. Nomikou and colleagues [24] were able to show that during a diaper change routine, there is a systematic coordination of eye gaze between mother and infant. Jaffe and colleagues [9, p. 1] defined coordination as interpersonal contingency or synchrony "such that each partner's behavior can be predicted from that of the other." Interestingly and contrary to Kaye's assumption for sucking bursts, there is neither a clear leader nor a follower in this behavior: Already at 3 months of age, infants are able to follow *and* to initiate turns via their eye gaze. Research suggests, however, that the contingencies can change due to the situational conditions and thus interactive (micro-)context in which they occur and as the infant develops. More specifically, Van Egeren and colleagues [25] have shown that contingencies are less pronounced when infants are held or become involved in object play. An additional factor playing a role in shaping the contingencies seems to be an infant's age: In the course of development, the contingency changes in terms of frequency, rapidness, and duration and acquires the form of an effective coupling: Although "mother and infant repeat each other's behavior less" when infants are 6 months old, the mutual exchange is rapid with less variability in timing [24, p. 290]. The authors concluded that gaze behavior seems to be conversational from early on in the sense that it regulates social interaction. We therefore propose that at this young age, infants are not only responding to actions addressed to them but are also initiating them. In fact, Goldstein and colleagues [26] demonstrated an association between 5-month-old infants' vocalizations and responses from caregivers that the authors interpreted as a learned social efficacy.

As a term comprising turn-taking behavior, *interpersonal synchrony* [27], [9] extends the notion of an exchange by emphasizing the rhythmical properties of mother–child coordination from birth onward [28]. According to Trevarthen [29], the temporal organization of multimodal exchange is at the core of social communication among humans. Provasi and colleagues [30] consider that rhythm is the fundamental aspect—this is why Gratier [31, p. 535] suggests extending the term "interactional synchrony" to cover temporal and vocal/prosodic coordination. For language acquisition, the manifold benefits of turn-taking can be summarized as the foundations of both conversational mechanisms and phonological development. Concerning the conversational mechanisms, Gratier [31] suggests that the function of this form of synchrony for turn-taking behavior is to predict the other's actions quite precisely. The partners involved can rely on and play with each other's expectations [31]—a key element in building up participatory communication skills [32, 19]. Accordingly, vocal turn-taking was found to provide structural context to a conversation, to facilitate attempts to mimic, and thus to bring about positive arousal [21] and bidirectional attachment [9]. For phonological development, turn-taking already yields changes in the quality of 3-month-olds' vocalizations [21] toward speech-like syllabic rather than non speech-like sounds. This is considered to be a milestone in speech development (e.g., [11]).

It appears that temporal coordination plays an important role in early communicative development—and not only for vocal coordination in particular but also for the coordination of communicative behaviors in general (also comprising nonverbal modalities). In this vein, based on results from a longitudinal study, Rohlfing and Nomikou [33] revealed that more organized dyadic interaction at 3 and 6 months is correlated with more advanced vocabulary development at 24 months. We can only speculate that the experience of interpersonal contingency in the form of turn-taking behaviors might result in children developing meanings that are shared without yet having the content of individual words at their disposal [34].

Taken together, results on turn-taking and interpersonal synchrony contribute to our increasing knowledge about how young children and their caregivers organize their interaction, how they take turns in their vocal behavior, and how this system drives language acquisition. Nonetheless, little is known about how interaction is organized across modalities. Because human behavior is sequential and distributed over many modalities – before taking a vocal turn, for example, a child will look at her partner [35] – it is reasonable to assume that there is a contingent exchange not only within but also across modalities. In fact, for infants at the age of 9 months, parents will initiate a verbal exchange but react with gestures [36] or even bodily movements [37] when the child responds. This exchange across modalities might be particularly productive for development, because one means (e.g., vocal behavior) can be substituted by another means (e.g., gesture/gaze). Moreover, the expected responses might change throughout development. Filipi [38] provides examples of how an infant might first show understanding of adjacency pairs in a nonverbal modality. For the summons–answer adjacency pair, she shows how at 9 months of age, a child will respond to a summons such as calling her name with a turn of the head and gaze toward the parent. In this case, the child is able to fulfill her role in the sequence without speech. At this age, the parent will accept this as an appropriate answer. This makes it possible to practice the organization of an exchange long before vocal behavior is in

place; or it can be particularly useful when verbal behavior is emerging. More specifically, a longitudinal study by Hilbrink and colleagues [39] revealed that children's timing of their turns slowed down toward the end of their first year of life, namely around 9 months—at precisely the moment when infants start to produce their first words. Because speech production is considered to be the bottleneck for the whole language system [40], it is possible that children compensate for this hurdle with multimodal turn-taking; that is, by reacting across modalities. Later in development when they have already started to produce some words, infants gaze at their mothers at the end of their turns, handing over the turn to their parent [16]. The latter patterns resemble what we know from adult conversation. Furthermore, in analyses of early peer interactions, Kidwell [41] has shown how young children use gaze shifts to initiate actions. For example, a gaze toward a caregiver to seek assistance initiates a sequence that requires a response by the caregiver. Even more interestingly, her analysis shows that children monitor gaze to elicit or even avoid a response from others, thus using gaze to structure and shape their interactions. Filipi writes: "The act of attending to a person [...] creates a place for a response" [38, p. 7].

II. Turn-taking across modalities

Human behavior is multimodal and takes advantage of various behavior possibilities. When asked a question, we can respond verbally, but we can also just gaze or use a gesture, shake our head, and so forth. In the following, we refer to this phenomenon as *multimodal turn-taking*. In conversation analysis, all these various behaviors can be considered as a sequence and interpreted qualitatively. However, methods for performing quantitative analyses of relationships across modalities are still, to our knowledge, rather scarce.

One method that is quite well-known in developmental research is Bakeman and Quera's [42] sequential analysis using GSEQ software. It assesses the extent to which one partner's behavior (e.g., the behavior of the child) is a function of the preceding behavior of the other partner (e.g., the mother) and vice versa. Lavelli and colleagues [43], for example, have recently used this method to assess maternal communicative modalities and child's conversational responsiveness to them. Thus, the method works across modalities (see also [25]). Central to this approach is the positioning of nominal data (states) in a sequence and the computation of transitional probabilities between states. Inherent to the method is the use of categorized behavior. For example, any verbal behavior of the children and mothers is differentiated in terms of whether it is an answer, a question, a repair, and so forth. Moreover, the first state in a sequence is defined a priori. However, one criticism is that because the method analyzes nominal data and not intervals, it is somewhat static in that it cannot capture the temporal relationships between the onsets and offsets of behaviors. This is disadvantageous when investigating such a phenomenon as turn-taking in which the exact timing of the onset of turn and switching pauses is key. In sum, sequential analysis excludes the interactional dynamics of turn-taking, even though it can be used to assess the multimodality of behavior [25], [43]. If a method is to capture the interactional dynamics of interactive formats (episodes or streams of sequences), it needs to be able to preserve the timing and also relate different modalities to each other.

In the following, we will define the problem (Section A) and indicate some ways to analyze it (Section B). In Section 3, we present some new solutions to studying multimodal turn-taking.

A. *Methodological challenges*

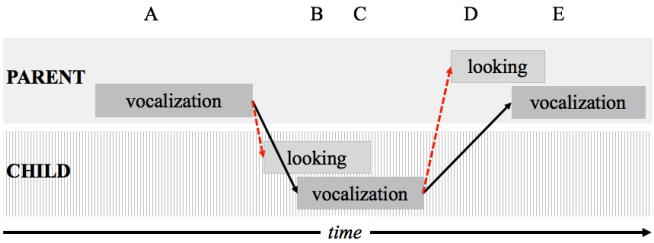

Figure 1: The data provides an example from a parent interacting with her/his child. Each participant is indicated by a layer. Within this interaction, some communicative behaviors in both partners can be observed that are depicted by a sequence of boxes (e.g., smiling, vocalization). The black arrows reflect a classical approach to turn-taking in which only vocal behavior is considered as relevant. Here, the problem is that such a unimodal approach omits other behaviors relevant to the communicative exchange. Multimodal turn-taking, however, involves any contingent behavior as indicated by the red and broken arrows.

Fig. 1 illustrates the type of data, with which we are dealing in multimodal turn-taking analysis. The data describes behaviors of participants (e.g., vocalizations, eye-gaze behavior, smiling) during a certain period of time. Each behavior is relevant to the communication and can be characterized by its modality and its temporal extension. More specifically, multimodal turn-taking data can be modeled in the form of a set of tuples (P_i, B_i, s_i, e_i), with P identifying the participant, B the behavior performed, s the time point when it starts and e the time point when it ends in an observation i.

On the basis of this kind of data, different questions can be addressed. For our purpose, Fig. 1 illustrates the main analytical problem of multimodal turn-taking analysis. The first question is how to capture a behavior that is organized as a (sometimes overlapping) sequence (A, B, C, D, E) but is spread across different modalities (vocalizations, eye-gaze behavior, smiling). The second question is how to capture the multimodal behavior that is spread across participants. In our example, both the parent and the child are co-constructing a sequence.

B. *Analysis methods*

Turn-taking data of this kind can be analyzed in various ways depending on the goals and purposes of an investigation. Generally, we can make the following distinctions with respect to data analysis:

- Confirmative versus exploratory: The data can be used to confirm or refute a certain hypothesis about turn-taking behavior, for example, that vocalization is paired with eye gaze. Alternatively, data analysis may serve a more exploratory purpose, that is, to help understand and gain insight into the data, generate hypotheses about sequences of a certain kind, and so forth.

- Descriptive versus inductive: A mere description of the observed data in the sense of aggregation, visualization, and so forth needs to be distinguished from the induction of models that generalize beyond those data.
- Explanatory versus predictive: The main purpose of a model could be to help to understand and explain the phenomenon of turn-taking, or, alternatively, to be able to predict and anticipate the behavior of persons in a given situational context.
- Global versus local: Analyses and models can be global in the sense of referring to the data and turn-taking behavior in their entirety, or rather local in the sense of analyzing only parts of the data or capturing only specific aspects of the behavior.

There are several reasons why turn-taking data are interesting from a methodological point of view:

- First, the data are of a *sequential nature* and combine *discrete* (type of event) with *continuous* information (time points and duration). Thus, we are dealing with heterogeneous temporal information that is more complex than, for example, standard time series data.
- Second, the data result from an interaction between two (or more) partners. Thus, as in a multiple time series analysis, we are dealing with *multiple information sources* that generate data in parallel.
- Third, the data are multimodal in nature. This means that even a single source may produce several events in parallel. Roughly speaking, this indicates that the dimensionality of the data is not fixed but may change over the course of time.

These points together reveal that two methodological issues are of particular importance: How should timing be interpreted? And how should individual states be defined? Whereas in the current article, we propose some answers to these questions, future research needs to broaden the scope of possible answers by taking into account the meanings of interpersonal situations (e.g., that a stroke of a gesture – which is the movement's peak when gesture is performed – is crucial for its meaningful occurrence).

In summary, we are dealing with sequential, mixed discrete-continuous, interactive, multimodal data. Various directions and approaches for analyzing such data are conceivable, although most methods will not be applicable in a straightforward way without further modification or extension:

- There is large body of literature on *probabilistic models* for sequential data, most notably Markov chains and hidden Markov models (HMMs). These can produce a better rendering of the multimodal interaction structure. However, in general, they are restricted to modeling the transition between *states* without capturing temporal information about time points or durations. In the case of turn-taking, states may correspond to the actions of the partners or, in HMMs, to mental states (the actions would then be considered as 'manifestations' resulting from the mental state). There are extensions of so-called stochastic automata that are closely related to Markov chains and capture temporal information about the time points of events. Nonetheless, in spite of their obvious potential, we are not aware of any application of models of this kind to turn-taking data so far.
- Because interaction between partners and interdependencies between their actions are of major interest in the analysis of turn-taking data, all sorts of *statistical correlation analyses* could be applied. Especially relevant here is correlation analysis in the context of time series that also captures time-shifted dependencies and time delays. Cross recurrence quantification analysis (CRQA) as presented in the next section falls into this category of methods that are mostly of a descriptive, exploratory nature.
- Going beyond standard statistics, *data mining* methods are specifically appropriate for dealing with complex, heterogeneous data. Especially interesting for turn-taking are frequent pattern mining techniques tailored for sequential, temporal data. Again, such methods are essentially of a descriptive, exploratory nature. In contrast to correlation analysis, they focus mostly on the extraction of *local* regularities in the data (see next section).

III. APPROACHES TO MULTIMODAL TURN-TAKING

In this section, we present first solutions to the problem as defined in section II A. First, we attempt to employ a dynamical time series analysis and present the CRQA as used within a single modality to showcase its time-preserving capabilities. Next, we use it to detect relationships among different modalities.

A. CRQA

Cross Recurrence Quantification Analysis (CRQA) is a method for analyzing time series based on nonlinear dynamic systems theory [44], [45], [46]. It builds upon the recurrence of similar states happening in two streams or time series. In the cognitive sciences, it has been applied to many types of behavioral streams – both continuous and categorical in nature – generated by two actors in order to establish (among other things) the degree of coupling between the two. Several possible measures relating to the dynamics of the two signals can generally be extracted, but also, crucially, a lag profile of the repetitions of same behaviors in the signals. For example, Richardson and Dale [47] used CRQA in a situation in which a speaker was talking about TV characters presented in some panels on a computer screen while listeners watched the same panels. They could observe a peak of recurrence at a lag of about 2 s—meaning that at any given moment, listeners tended to fixate the same panels that the speaker had fixated 2 s before. Moreover, the peak was present only if listeners actually demonstrated comprehension of what was being said [47]. Hence, this kind of analysis is able to capture whether a specific kind of behavior (e.g., fixating the same region of the computer screen) by one of the actors in the interaction is mirrored consistently at any given lag by the same behavior in the second actor—a modality-specific behavioral coupling.

In the specific context we are considering here, that is, the emerging turn-taking behavior of mothers and their 3-month-

old infants, CRQA has been used to unveil the fine time tuning of behaviors in the two interaction participants within specific modalities. Nomikou et al.'s [24] study considered only gazing behavior (defined as gaze at interaction partner's face), and used CRQA to confirm a tight synchronization of gazing between infants and their mothers and also show how this changes developmentally from the ages of 3, to 6, to 8 months. Similarly, Leonardi et al. [48] applied the same method to the vocalization of behavioral streams and found a systematic lagging of mother's vocalizations after infants' vocal production together with an active avoidance of synchronous vocal behavior by the two actors; that is, first signs of vocal turn-taking.

CRQA has therefore usually focused on within-modality coupling; that is, the recurrence of the same behavior in two behavioral streams of the same kind. This is probably due to the fact that it is easier and logically more consistent to map similar behaviors (or events in a behavioral stream) against themselves. For example, in the case of fixation to areas on a computer screen, it makes sense to count as recurrent the behavioral events of fixating in the same area of the screen. When recurrence cumulates at particular lags, this would indicate a consistent matching of 'same-behavior' (i.e., fixations in the same area) at specific delays, hence abstracting from which area was actually fixated at every given moment. How behaviors in different modalities could then be mapped across themselves to count as recurrent is, at first, slightly less intuitive.

We achieved this by simplifying the behavioral coding in the various modalities to a series of binary codes: presence or absence of a given behavior. Using the same corpus as in [24] and [48], we asked whether gazing behavior could be related to vocalizations; in other words, whether there is a systematic relationship in time between the production of any kind of vocalization in one of the actors and gazing in the other.

The methodological solution we adopted was to code the occurrence of a vocalization and the occurrence of a gaze-at-face behavior in the same way (e.g., with the same value 1) and to code the absence of such a behavior with a null value (0). In this way, recurrences in the analysis would indicate a match of behaviors across modalities.

If we want to obtain a complete picture of the multimodal turn-taking between the two actors (mother and infant), one first challenge with this kind of approach is the need to run as many analyses as there are pairwise combinations of the possible behavioral modalities that we intend to explore. For example, in the case of two modalities as proposed above, we need to analyze the coupling (i.e., the recurrent behavior) of mother's gaze and infant's vocalizations as well as the coupling of infant's gaze and mother's vocalizations, adding to them also the analyses run within the same modalities between the two actors (i.e., mother's and infant's vocalizations on one side and mother's and infant's gaze on the other). We can well imagine that the possible interrelation of many behavioral streams at once could quite rapidly reach a level on which a global interpretation of the analyses would become intractable.

An additional point to consider in this kind of multimodal analysis is that in order to emerge as an established temporal pattern, the relationship of the two different behavioral streams has to be not only consistent across the sample but also routinized; that is, it should have a higher probability of repeating itself over the course of the interaction as well as across different dyads.

To show what the possible outcomes of such analyses could look like, we took 16 dyads of the corpus described in [24] and [48], in which we analyzed unimodal coupling for both gaze and vocalizations, and we extracted an additional behavioral stream: smiles of mother and infant during the interaction. All the behavioral streams were coded from video-recorded diaper-changing sessions by trained coders who annotated the starting and ending times of such behaviors (i.e., gaze-at-face, vocalization, and smile for both actors) in dyads with 3-month-old children.

When preparing the data for the analysis, we turned the annotations into binary time series sampled at a 10 Hz sampling rate. Because the occurrence of all behaviors was coded in the same way by using the same numerical code for all of them, we proceeded by analyzing the time-lagged recurrence profiles extracted from the cross-recurrence plots of every combination of behaviors and actors manifesting them. Fig. 2 presents the averaged recurrence profile of the multimodal coupling of gaze and smile.

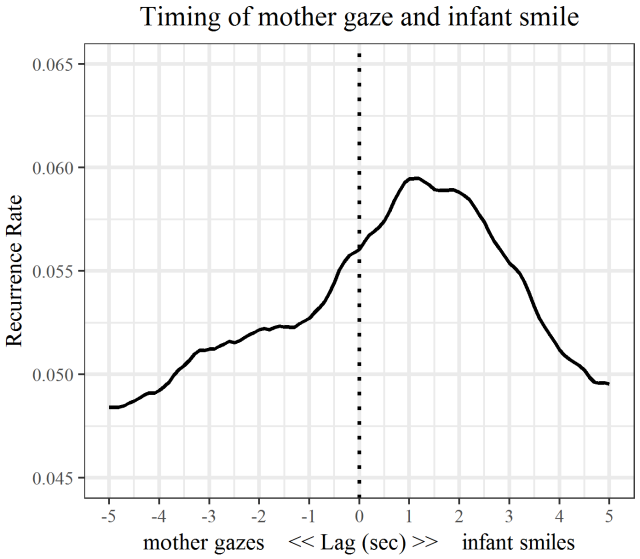

Figure 2. Averaged recurrence profile of the behavioral streams of mothers' gaze-at-face and infants' smile across 16 dyads. At the central dotted line, at lag 0, the level of recurrence corresponds to the amount of synchronous (i.e., manifested at the very same moment in time during the interaction) appearance of the two streams. On the right, we have the amount of recurrence of infant's behavior at different lags following the mother's behavior (here the infant's smile following mother's gaze-at-face). The left side shows a similar relationship but reversed among the actors (here the mother's gaze-at-face following the infant's smile).

In Fig. 2, the recurrence lag profile indicates an increased probability of an infant smile in the first two seconds after the mother gazes at the infant's face—a finding also reported in Szufnarowska and Rohlfing [49] who used traditional methods (which extensively analyzed the conditions under which a smile of the infant would occur). However, in [49], the reciprocity of the behavior (i.e., the probability of mothers smiling in return,

see Fig. 3) was not taken into account. Thus, by applying CRQA, we could also explore whether the reverse is true by running a cross-recurrence analysis of the infant gaze behavioral stream with the mother smile. The outcome of this analysis is shown in Fig. 3.

In Fig. 3, we can see how the probability of mothers smiling in response to their child gazing straight at them seems to increase within the first 2 s after the beginning of the infant's behavior (gaze at mother's face). This shows as a peak on the left side of the graph.

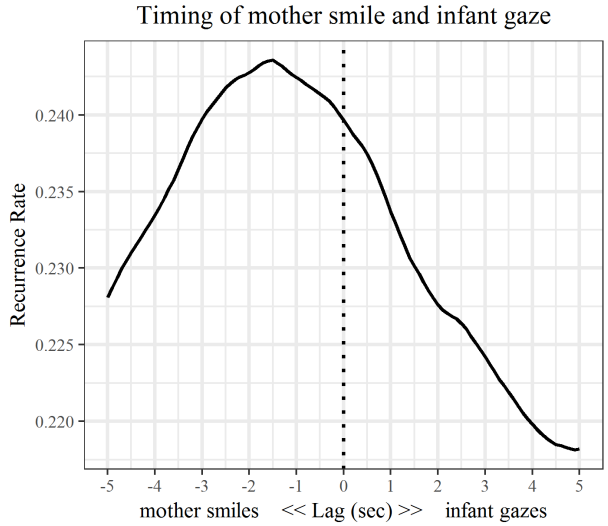

Figure 3. Averaged recurrence profile of the coupling across modalities in mother–infant dyads. Mothers' smiles are matched with infants' gaze-at-face. The right side of the graph depicts infant's gazes following mother's smiles, whereas the left side depicts mother's smiles following infant gazes.

Similar combinations of recurrence profiles can be computed for vocalizations and smiles as well as for gaze and vocalizations, again in both directions, and obviously for intramodal turn-taking: vocalizations coupling, gaze, and smiles coupling across the actors.

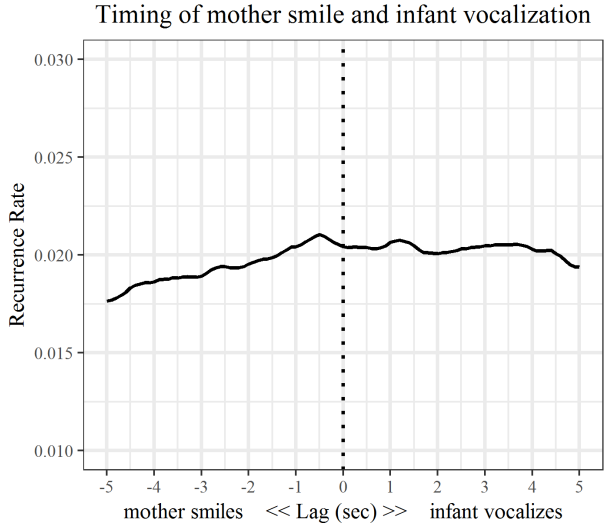

Figure 4. Averaged recurrence profile of the coupling across modalities in mother–infant dyads. Mothers' smiles are matched with infants' vocalizations. The right side of the graph depicts recurrence of infant's vocalizations following mother's smiles, whereas the left side depicts the recurrence of mother's smiles following infant vocalizations.

Fig. 4 presents an example in which it seems that no particular pattern in the turn-taking behavior of mothers and infants is present. It relates to the coupling of mothers' smiles and infants' vocalizations. In Fig. 4, the cross-recurrence does not seem to gather consistently at any particular lag in the time-dependent relationship of the two behaviors under consideration.

Vice versa, when looking at the analyzed coupling between mothers' vocalizations and infants' smiles (Fig. 5), there seems to be some trend for recurrence to pile up on the right-hand side of the recurrence profiles: on the side of the infant' reactions about 1 s after their mothers' vocalization. This trend shows up more clearly at later ages (data not shown).

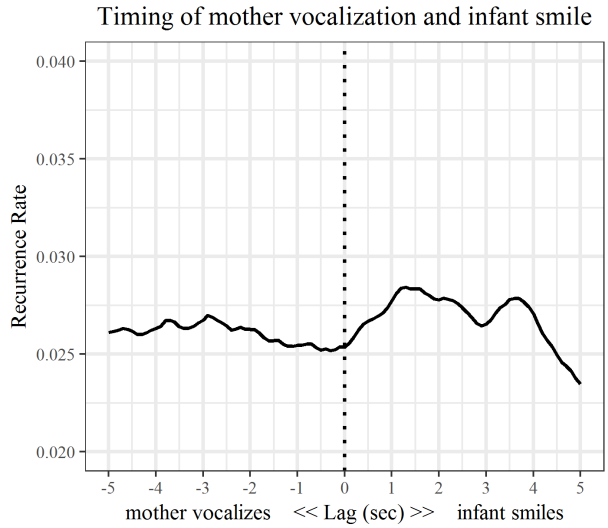

Figure 5. Averaged recurrence profile of the coupling across modalities in mother–infant dyads. Mothers' vocalizations are matched with infants' smiles. As before, the right side of the graph depicts the degree of recurrence of infant's behavior following mother's vocalizations, whereas the left side depicts mother's vocalizations following infant smiles.

In the above analyses, we demonstrated the general feasibility of this multimodal turn-taking analysis by using CRQA and its associated recurrence profiles in a predefined lag window. However, the results shown above need to be analyzed for their statistical significance before the increased probability of a behavioral coupling (manifested by a peak in the recurrence profile) in two behavioral streams can be confirmed. Moreover, the several combinations of behavioral coupling given by only a few analyzed streams quickly give rise to many possible analyses that can be difficult to synthesize in any unified account. Nonetheless, it is likely that the best results will be achieved by combining the probability-based methods – which give an overall sequence structure – with CRQA, which, as we have demonstrated above, can complement the picture with information about the exact timing. We also need to ask whether CRQA assesses turn-taking per se or whether it is limited to capturing a coupling or synchronization. Such questions definitely stimulate further a discussion over whether these are the same phenomena or whether they are just related.

In general, as a proof of the concept, the method presented here proved to be a promising way to approach multimodal turn-taking in human behavior.

B. Frequent Pattern Mining

In this analysis, we considered the same 16 dyads. As in the data presented within the CRQA analysis, infants were 3 months of age. We used a specific data mining approach to shed light on emerging patterns.

Data mining is a relatively young research field at the intersection of computer science and statistics. The main task of data mining is to search data for potentially interesting patterns, whereby the meaning of 'interesting' may depend on the application and the purpose a pattern is being used for. Quite often, interestingness is connected to the *frequency* of occurrence: A pattern is considered interesting if the number of its occurrences in the data strongly deviates from what one would expect on average. A *frequent pattern* is thus a pattern simply observed much more often than others, and the problem of discovering such patterns is called *frequent pattern mining* [50]. The other extreme is outliers and *exceptional patterns* that deviate from the norm and occur rarely in the data. Finding such patterns might be of interest, too, and is called *exception mining* [51]. Moreover, other forms of data mining also exist. In *contrast mining*, for example, the interest is in finding patterns that distinguish different subpopulations within the data—for example, ones that are significantly more frequent in one subpopulation than in another one [52].

Patterns are mostly of a *local* nature, pertaining only to a (small) part of the data but not describing the data as a whole. Moreover, data mining is an *exploratory* endeavor, and patterns discovered in a data set are usually interpreted in a *descriptive* way. This is in sharp contrast to inferential statistics and confirmative statistical analysis in which the data are used to confirm or reject a pre-specified hypothesis (that has been generated independently from the data). Instead, data mining typically serves the purpose of *generating* hypotheses that then need to be validated (or refuted) in a subsequent step. Typically, a large portion of the patterns extracted by an algorithm turns out to be uninteresting, either because the patterns merely reproduce dependencies the domain expert is already aware of, or represent artefacts in the data that cannot be generalized.

The type of patterns considered and the criteria used to assess their interestingness depend strongly on the nature of the data. It makes a big difference whether, for example, data are binary, categorical, or numerical, and whether a single observation is described in terms of a subset as in so-called itemset mining or as a sequence as in sequential pattern mining [53]. Likewise, the type of data will have a strong influence on the algorithms that are used to extract the presumably most interesting patterns. From an algorithmic point of view, the key challenge is to design algorithms (including suitable data structures) that extract patterns efficiently and can avoid any excessive time or space complexity. This is challenging due to the sheer size of the search space: Except for trivial cases, the number of candidate patterns is huge, and often grows exponentially in certain characteristics of the problem instance (e.g., the number of items in itemset mining).

Especially relevant for the problem of turn-taking is the analysis of sequential data in which individual data items (social actions, types of behavior, etc.) have a temporal order. Thus, it is possible to characterize the temporal relationship between events. For example, events can overlap, or one event can occur after another one. Correspondingly, sequential patterns are expressed in terms of these types of relationships. Temporal data can be seen as a specific type of sequential data in which events typically have a starting point, a duration, and an endpoint. Thus, temporal data mix two types of information: discrete (a certain type of event has occurred) and numerical (its location and extension in time).

As an example, consider the interval data shown in Fig. 6. Each interval represents the start, duration, and end of the occurrence of a certain type of event (indicated by letters A, B, C). Such intervals can overlap; that is, several events can occur in parallel. An interval sequence of this kind can be divided into another unique sequence of consecutive intervals (indicated by the numbers 1 to 13 in the figure) such that a *set of events* (such as A, B, C) occurs in each of these intervals.

An example of a pattern that we subsequently consider to be an important special case could then be a rule of the form S → T in which S and T are subsets of events. This could be interpreted as follows: If the events in S occur (simultaneously), then the events in T occur shortly after. Such rules are sometimes called 'association rules' because they establish a causal association between S and T, suggesting that S triggers the occurrence of T.

To evaluate such a pattern, one can count the number P of positive examples in the data set (i.e., occurrences of S followed by T) and the number N of negative examples (occurrences of S not being followed by T). A typical measure for assessing a candidate rule derived from these numbers is the *confidence* P/(N+P) that can be interpreted as an (estimated) conditional probability that T occurs (shortly) after S has occurred. Thus, a high confidence (together with a sufficiently large number P of positive examples) suggests a strong dependency between S and T. In this regard, it is important to note that a confidence close to 1 is not necessarily required. Instead, a pattern with a much lower confidence could already be interesting. In fact, the interestingness of a rule depends strongly on the 'default' (prior) probability of T, and how this probability is increased by the occurrence of S. For example, a rule S → T with confidence 0.3, despite looking low at first sight, could be considered interesting if T occurs only very rarely in general (say, with a probability of 0.1), because this means that the occurrence of S significantly increases the probability of the occurrence of T.

In our data mining algorithm, which we implemented as a Java program, we used the following conditions for a positive example: The earliest starting point of T is the starting point of S (one could also include some delay that differs for the adult in comparison to the infant as in [54]). Moreover, the latest starting point of T is the endpoint of S. An occurrence of S that is not overlapped or succeeded directly by T is counted as a negative example. Our algorithm systematically searches all possible candidate patterns up to a certain size (not more than five events/behaviors in S) and computes their support through simple counting.

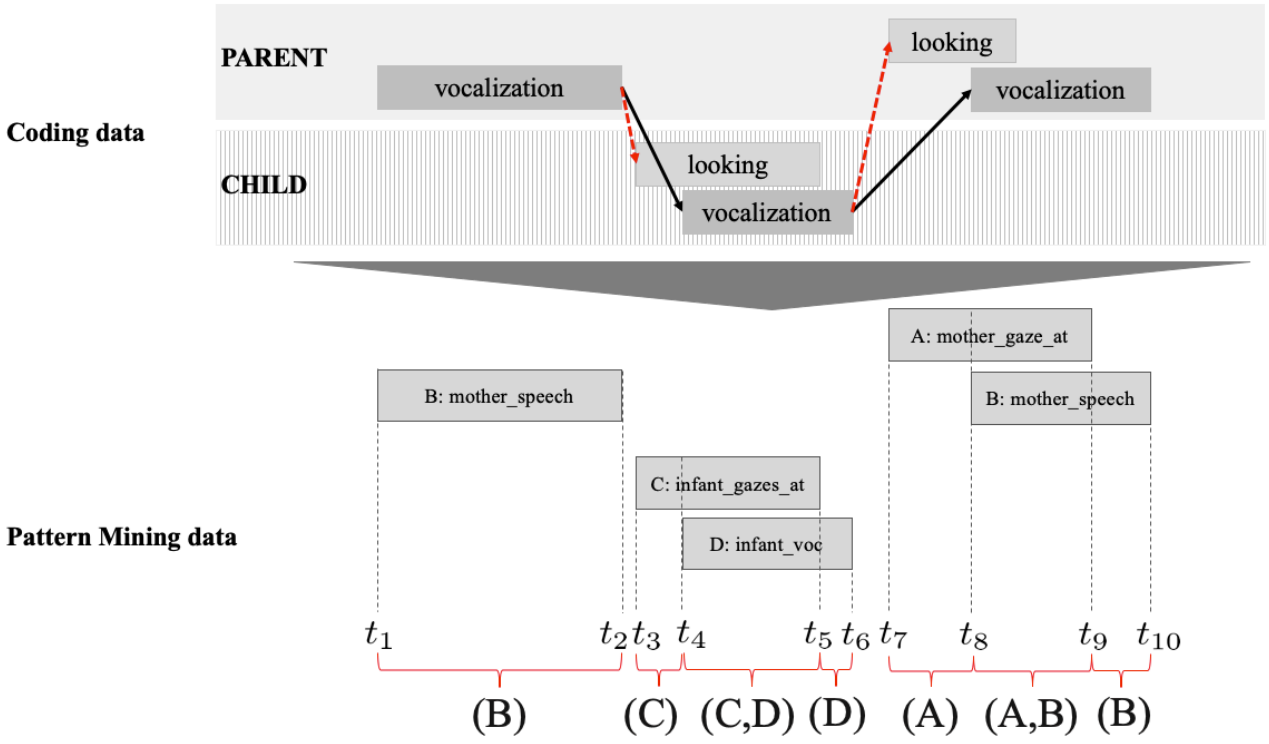

Figure 6. Interval data that is transformed from the coding data (top) to sequences (on the bottom).

The following analysis is based on data about the behaviors of the partners presented in Table 1. As can be seen in Table 1, in comparison to CRQA, we were able to differentiate between a direct gaze toward the partner within an interaction and a gaze toward an object. In addition, we were also able to consider patterns of several behaviors being performed simultaneously by a single person.

TABLE 1
BEHAVIORS INPUTTED FOR THE ANALYSIS

mother's	infant's
mother_speech	infant_voc (vocalizations of the infant)
mother_smile	infant_smile
mother_gaze_at (gazing at her baby)	infant_gaze_at (gazing at her/his mother)
mother_gaze_away (gazing away from the infant but not at an object)	infant_gaze_away (gazing away from her mother but not at objects)
mother_gaze_at_objects (gazing at objects)	infant_gaze_at_object (gazing at an object)

In the following, we highlight some patterns. This selection was based initially on the following criteria: high confidence, verbal behavior of the mother, verbal behavior of the infant, and comparisons to CRQA. Premises are indicated in brackets, and conclusions are followed by information about the probability of the occurrence, stating, for example, that when an infant is gazing at the mother, the mother will gaze at the infant with a probability of 50% as in Pattern (1).

The first two patterns with the highest confidence are quite trivial, suggesting that when the mother looks at the object, she will then look at the infant and vice versa. This is due to the nature of the activity in which she is engaging (e.g., changing diapers), with her attention (i.e., gaze) being split between baby and objects that are relevant for this activity. Looking at the first patterns with the highest confidence, it appears that the infant – and more specifically, the infant's gaze at the mother – is dominating the premises. Nomikou et al. [55] confirmed this when they found that at the age of 3 months, the infant is gazing at the mother during a major part of the interaction. However, due to the activity the mother is performing, there is a probability of (only) 50% that the mother will look back when the infant is looking (Pattern 1). Interestingly, infant's gaze as a conclusion to maternal gaze occurs on a lower confidence level of 19% (2).

(1) `(infant_gaze_at)` → `(mother_gaze_at): 0.50`
(2) `(mother_gaze_at)` → `(infant_gaze_at): 0.19`

Some patterns of maternal vocal behavior also become apparent. First, (3) and (4) present two patterns with a quite high confidence that suggest an intrapersonal synchronization of smiling and speaking (3) that can also be paired with gazing at the infant (4). In contrast, the pattern in (5) suggests an interpersonal sequence of mother gazing at her infant together

with infant gazing back, resulting in vocal behavior by the mother.

(3) (mother_smile) → (mother_speech): 0.70

(4) (mother_gaze_at, mother_smile) → (mother_speech): 0.65

(5) (infant_gaze_at, mother_gaze_at) → (mother_speech): 0.66

Pattern (5) indicates a co-constructive framing of behavior, with mother's gaze, paired together with infant's gaze, being followed by mother speaking. Thus, the complete loop of verbal interaction here seems to depend on mutual gaze—a phenomenon considered recently in Nomikou et al. [56]. Probably, this framing will become less prevalent as the child gains in interactional experience. The co-constructive framing also becomes visible when considering the child's vocal behavior (see later Patterns 15 and 16).

With respect to the effects of maternal vocal behavior, we see that it is followed by infant's gaze toward the mother with a confidence of 13%, which seems rather low. Interestingly, when considered in a specific context, namely when the infant is gazing at an object, confidence is higher (6). The same is true for infant gazing away.

(6) (infant_gaze_at_object, mother_speech) → (infant_gaze_at): 0.30

(7) (infant_gaze_away, mother_speech) → (infant_gaze_at): 0.23

These patterns underscore the attention-regulating (ostensive) character (and its effect) of maternal speech [57], but also the need for this signal to occur in a rich context if the child is attending to something else.

So far, the analysis has revealed a number of potential responsivity patterns in terms of the dyadic nature of turn-taking; that is, the fact that the two partners are responding contingently to each other. However, a more advanced form of interaction is to coordinate toward an entity in the world—a pattern that is typical for a triadic interaction. The first such pattern, with 42% confidence is presented in (8).

(8) (infant_gaze_at_object) → (mother_gaze_at_object): 0.42

(9) (mother_gaze_at_object) → (infant_gaze_at_object): 0.17

We see that the infant's behavior is followed by the mother (8) on a higher confidence level than the mother's behavior is followed by the infant (9). This occurrence suggests that in the triadic interaction, the child might be the driving force in establishing the pattern, as suggested by [50].

In CRQA, we took infant's and mother's smile into consideration. Results showed a high probability of infant's smile occurring when it is preceded by mother's gaze at the infant. We found this pattern at a confidence level of 35% (10). The highest confidence level for infant's smile was preceded by both, mother and child gazing away (11).

(10) (mother_smile) → (infant_smile): 0.35

(11) (infant_gaze_away, mother_gaze_away) → (infant_smile): 0.38

For the infant's smile to occur, it thus seems necessary to gaze away right before. One has to bear in mind, however, that our subjects are 3 months old, and the pattern might reflect the fact that interaction is progressing while the child 'works' on producing the smile triggered by an earlier cue. In fact, in a recent approach to calculating the significance among the patterns, this particular pattern seemed to be nonsignificant [54]. In this vein, Fig. 3 shows that infants' smiles come roughly 1 s after mothers' gaze. If other behaviors are interspersed within this second, frequent mining will 'catch' them and assign them to the smile, thus missing the delayed vocalization–smile contingency. This is clearly a limitation of the frequent pattern mining method we applied. However, it is known from RQA that there are some analyses in which it might be better to apply a somewhat coarser resolution to reveal higher-order structures. Ruland [54] has demonstrated that delays can be implemented in the pattern mining method, and we will address this possibility in the Discussion section.

For mother's smile, the CRQA analysis revealed that it is coupled with infant gazing toward the mother. The pattern mining analysis can confirm that mother's smile is followed by infant's gaze (12) and vice versa (13), with the difference in confidence matching the CRQA recurrent rate values for the two orders of these events.

(12) (mother_smile) → (infant_gaze_at): 0.23

(13) (infant_gaze_at) → (mother_smile): 0.53

Finally, infants' vocalizations were considered in CRQA. Between maternal smile and infant's vocalization, the cross-recurrence did not seem to consistently yield any time-dependent relationship of the two behaviors. Applying pattern mining, we can see in conclusion that infants' vocalizations occur after various behaviors at the confidence level of 48% and less (see Patterns 14–16). However, in the future, we need to further investigate what the markers of significance are when using this method. We also see that this pattern contains infants' behavior as well as the change in the gaze of the mother: As an antecedent to infants' vocalization, a complex co-occurrence of infant smiling and maternal orientation away is apparent (16). In other words, patterns involving infants' vocalization include maternal orientation away from the dyad with quite high confidence. This speaks to the co-constructive complexity of social behaviors within which infants' vocalizations are embedded.

(14) (infant_gaze_at) → (infant_voc): 0.47

(15) (infant_gaze_at, infant_smile) → (infant_voc): 0.41

(16) (infant_smile, mother_gaze_away, mother_speech) → (infant_voc): 0.48

It is likely that with phonological development (e.g., at the child's age of 6 months), more patterns will be comprised of vocal behavior—a hypothesis that has to be tested in future analyses of longitudinal data.

Taken together – with CRQA and frequent pattern mining analyses – we aimed to provide some support and initial results for the proposition that human interactive behavior has a discernible sequential organization. The results support the concept of pragmatic frames recently discussed by Rohlfing and colleagues [59]. Accordingly, verbal and nonverbal behavior are co-constructed by the interaction partners and form a multimodal pattern that is at the core of communicative exchange. Above, we were able to present such multimodal patterns emerging between infants as young as 3 months and their mothers. In future analyses, we will follow the emergence of these patterns over the child's development by taking longitudinal data into account. In addition, we will consider other patterns of interactions such as the development of joint attention (i.e., how infants' gaze-following behavior is framed by what kind of maternal behavior) or the development of patterning of vocalizations. Following Nomikou and colleagues [55], we argue that some patterns will educate and reward the child's behavior.

IV. Discussion

We started our paper by reviewing literature that mostly considers turn-taking to be unimodal. Although there has been some tradition of descriptive, qualitative single-case analyses that have revealed the multimodality of turn-taking, generalization to bigger data sets in the form of quantitative analyses is still very scarce (but see [25], [43]).

According to theoretical positions emphasizing that communication is organized by the interaction partners jointly (e.g., [12], [59], [60]), we then defined the challenge of assessing human sequential behavior that is (a) spread across different modalities and (b) co-constructed with a partner. As a first solution to this challenge, we presented analyses of a corpus consisting of multimodal codings of 16 dyadic interactions. In these interactions, mothers interacted with their 3-month-old children during a diaper change (see, e.g., [61]). As a method accounting for various modalities, we first applied CRQA and were able to show that infant's smile occurs when it is preceded by mother gazing at her or him. The power of this method is that we can determine the time-dependent course of multimodal behavioral streams across the different actors engaging in an interaction in great detail and with good visualization.

As a second approach to the multimodal turn-taking problem, we applied Frequent Pattern Mining. The most frequent patterns with the highest levels of confidence revealed the rich structuring of turns provided by mothers, confirming existing research on self-synchrony and the multimodal framing of turns with ostensive signals (e.g., [61]). Furthermore, the analyses revealed that mothers' vocalizations were followed by infant gaze shifts toward the mother's face as well as by infant smiles; that there are direct responses of the partners to each other (e.g., Patterns 10 and 12); and that some behaviors are embedded into an extensive pattern (e.g., in Patterns 11 and 16). Thus, we see first indications for the proposition that some turn-taking is multimodally more complex than other turn-taking. Finally, our findings confirm the CRQA analyses [24] on the coordination of gaze toward the partner's face (pattern 1 and 2) and reveal different confidence values (50% of mother gazing at infant after the infant gazed at her and 19% for infant following the gaze of the mother). [54] was able to show that both patterns are significant when tested against a null hypothesis, thereby replicating previous findings [24] while taking a different methodological approach.

The two approaches presented here clearly show how the inclusion of multiple modalities is now becoming increasingly attainable in quantitative terms and how this enables a more elaborate description of young infants' communicative contributions as well as the complexity of the patterns in which they are embedded. This is critical for analyses of early interactions. By allowing for contingencies, dependencies, and complex patterns to emerge from multiple resources, analyses using the principles applied in the approaches above can show how subtle turn-taking skills – previously attributed to older infants – can now be investigated in younger infants as well. Future research can focus on how multimodality leverages unimodal behavior.

One of the differences between the two approaches is that CRQA seems to be more hypothesis-driven, because a decision needs to be made beforehand about which pairs of modalities to map against each other. Moreover, careful pairings of the coded categories across modalities need to be considered. Frequent Pattern Mining is a more exploratory method, because the premises and conclusions emerge as a result of the analysis. This is why for the present, exploratory stage, the methods can be seen as complementary, with frequent pattern mining providing the hypotheses regarding concrete modalities to be further investigated, and the CRQA supplying detailed picture of time-dependency of the selected modalities.

Future research should apply visualization techniques displaying the results of pattern mining in combination with statistical tests, because currently, we have simply picked some patterns from a large list. Another issue concerns the basis for the analyzed patterns: Whereas currently, contingency between behaviors is based mainly on time closeness, further research will find out about what level of coarseness is needed to tap into higher-order structures. Ruland [54] has explored the possibility of applying some time delays between the intervals for the pattern mining analysis. He justifies his approach by saying that reaction times differ between infants and adults. However, while it seems reasonable to consider delays in co-constructing sequences, we currently do not know enough about infants' reaction times within natural interactions to choose the appropriate interval.

In addition to the aforementioned limitations, Jaffe and colleagues [9] already pointed out already that interpersonal coordination as a more global form of turn-taking can serve different purposes while varying in its degree of coordination. More specifically, whereas a midrange degree of coordination

was considered to be optimal for the development of emotional attunement, high coordination captured a way of interacting that supports cognitive development [9, p. 108]. This work alerts us to the current limits of our approach focusing on the coordination per se and disregarding the fact that its degrees might serve different areas of development.

Developmental research has a strong interest in analyses of longitudinal data. This perspective brings new challenges to studying the emergence of patterns. From developmental research, we know that sequences in parent–child interaction are not fixed, and that they change during the course of a child's development. For example, Filipi [38] has shown that, while early in development, a behavior serves as a part of communication, later in development, the same behavior might initiate a repair sequence from the parent. Similarly, regarding the use of gaze in turn-taking by children themselves, D'Odorico, Cassiba, and Salerni [62] found that infants use gaze and vocal behavior in different ways at different ages: At about 10 months, infants attracted their mothers' attention through gaze at the opening of turn-taking sequences and then vocalized. Future research should therefore investigate how patterns of interaction change. Computational methods that can detect a transformation of a pattern will help us to address this issue.

ACKNOWLEDGMENTS

We would like to thank all participating children and parents, our great research assistants (Monique Koke, Bettina Wagner, Alicja Radkowska and Urszula Kalinowska-Drozd) and Sascha Henzgen for his work with the pattern mining scripts. We are indepted to the two reviewers for their constructive comments. This work was made possible within Beethoven project EASE funded by the NCN-DFG collaboration (RO 2443/5-1 for KR and UMO-2014/15/G/HS1/04536 for JRL).

REFERENCES

[1] T. Stivers, N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoyman, F. Rossano, J. P. de Ruiter, K. Yoon, and S. C. Levinson, "Universals and cultural variation in turn-taking in conversation," *Proceedings of the National Academy of Sciences,* vol. 106(26), 10587–10592, 2009.

[2] S. C. Levinson, "Turn-taking in human communication—origins and implications for language processing," *Trends in Cognitive Science*, vol. 20, 6–14, 2016.

[3] C. Demuth, "Handling power-asymmetry in interactions with infants: A comparative socio-cultural perspective," *Interaction Studies*, vol. 14(2), 212–239, 2013.

[4] D. N. Stern, J. Jaffe, B. Beebe, and S. L. Bennett, "Vocalizing in unison and in alternation: Two modes of communication within the mother-infant dyad," *Annals of the New York Academy of Sciences,* vol. 263(1), 89–100, 1975.

[5] H. Sacks, E. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking in conversation," *Language,* vol. 50, 696–735, 1974.

[6] E. A. Schegloff, "Overlapping talk and the organization of turn-taking for conversation," *Lang. Soc.*, vol. 29, 1–63, 2000.

[7] L. Henry, A. J. F. K. Craig, A. Z. Lemasson, and M. Hausberger, „Social coordination in animal vocal interactions. Is there any evidence of turn-taking? The starling as an animal model," *Frontiers in Psychology*, vol. 6, 1416, 2015.

[8] M. Stevanovic, and S. Peräkylä, "Experience sharing, emotional reciprocity, and turn-taking," *Frontiers in Psychology*, vol. 6, 450, 2015.

[9] J. Jaffe, B. Beebe, S. Feldstein, C. L. Crown, M. D. Jasnow, P. Rochat, and D. N. Stern, "Rhythms of dialogue in infancy: Coordinated timing in development," *Monographs of the Society for Research in Child Development*, vol. 66(2), i-149, 2001.

[10] M. Gratier, E. Devouche, B. Guellai, R. Infanti, E. Yilmaz, & E. Parlato-Oliveira, "Early development of turn-taking in vocal interaction between mothers and infants," *Frontiers in Psychology*, vol. 6, 1167, 2015.

[11] N. Masataka, *The onset of language.* Cambridge: CUP, 2003.

[12] J. S. Bruner, "The ontogenesis of speech acts," *Journal of Child Language*, vol. 2, 1–19, 1975.

[13] M. C. Bateson, "Mother-infant exchanges: the epigenesis of conversational interaction," Annals of the New York Academy of Sciences, vol. 263, 101–113, 1975.

[14] M. Argyle, *Social interaction.* New Brunswick / London: Transaction Publishers, 1973.

[15] C. Goodwin, "Restarts, Pauses, and the Achievement of a State of Mutual Gaze at Turn-Beginning", in *Sociological inquiry*, vol. 50(3-4), 272–302, 1980.

[16] D. R. Rutter, and K. Durkin, "Turn-taking in mother–infant interaction: An examination of vocalizations and gaze", *Developmental Psychology*, vol. 23(1), 54–61, 1987.

[17] K. Kaye, "Toward the origin of dialogue," in *Studies in mother- infant interaction*, H. R. Schaffer, Ed. London: Academic Press, 1977, pp. 89–119.

[18] K. Kaye, and A. J. Wells, "Mothers' jiggling and the burst—pause pattern in neonatal feeding," *Infant Behav. Dev.*, vol. 3, 29–46, 1980.

[19] K. Kaye, and R. Charney, "How mothers maintain "dialogue" with two-year-olds," in *The social foundations of language and thought: essays in honor of Jerome S. Bruner*, D. Olson, Ed. New York: Norton, 211–230, 1980.

[20] G. Gergely, and J. S. Watson, "Early socio-emotional development: contingency perception and the social-biofeedback model," *Early Soc. Cogn.*, vol. 60, 101–136, 1999.

[21] K. Bloom, A. Russell, and K. Wassenberg, "Turn taking affects the quality of infant vocalizations," *Journal of Child Language*, vol. 14(2), 211–227, 1987.

[22] Y. Nagai, Y., "Mechanism for cognitive development," in *Cognitive Neuroscience Robotics A*. Springer Japan, 2016, pp. 51–72.

[23] T. Striano, A. Henning, and D. Stahl, "Sensitivity to social contingencies between 1 and 3 months of age," *Developmental Science*, vol. 8, 509–518, 2005.

[24] I. Nomikou, J. Leonardi, K. J. Rohlfing, and J. Rączaszek-Leonardi, "Constructing interaction: The development of gaze dynamics," *Infant and Child Development*, vol. 25, 277–295, 2016.

[25] L. A. Van Egeren, M. S. Barratt, and M. A. Roach, "Mother–infant responsiveness: Timing, mutual regulation, and interactional context," *Developmental Psychology*, vol. 37(5), 684–697, 2001.

[26] M. H. Goldstein, J. A. Schwade, and M. H. Bornstein, "The value of vocalizing: Five-month-old infants associate their own noncry vocalizations with responses from caregivers," *Child Development*, vol. 80(3), 636–644, 2009.

[27] W. S. Condon, and Ogston, W. D., "Speech and body motion synchrony of the speaker-hearer," in *Perception of language.* D. L. Horton, and J. J. Jenkins Eds. Columbus, OH: Charles E. Merill, 1971, pp. 150–184.

[28] W. Condon, and L. Sander, "Neonate movement is synchronized with adult speech: interactional participation and language acquisition," *Science*, vol. 183, 99–101, 1974.

[29] C. Trevarthen, "Communication and cooperation in early infancy: a description of primary intersubjectivity," in *Before Speech: The Beginning of Human Communication,* M. Bullowa, Ed. London: Cambridge University Press, 1979, pp. 321–347.

[30] J. Provasi, D. I. Anderson, and M. Barbu-Roth, "Rhythm perception, production, and synchronization during the perinatal period," *Frontiers in Psychology,* vol. 5, 1048, 2014.

[31] M. Gratier, "Expressive timing and interactional synchrony between mothers and infants: Cultural similarities, cultural differences, and the immigration experience," *Cognitive Development*, vol. 18(4), 533–554, 2003.

[32] J. S. Bruner, *Child's Talk: Learning to Use Language.* New York, NY: Norton, 1983.

[33] K. J. Rohlfing, and I. Nomikou, "Intermodal synchrony – as a form of maternal responsiveness – is associated with language development," *Language, Interaction and Acquisition,* vol. 5(1), 117–136, 2014.

[34] C. Trevarthen, "What is it like to be a person who knows nothing? Defining the active intersubjective mind of a newborn human being," *Infant and Child Development,* vol. 20(1), 119–135, 2011.
[35] F. Franco, and G. Butterworth, "Pointing and social awareness: Declaring and requesting in the second year," *Journal of child language*, vol. 23(2), 307–336, 1996.
[36] C. Lüke, U. Ritterfeld, A. Grimminger, U. Liszkowski, and K. J. Rohlfing, "Development of pointing gestures in children with typical and delayed language acquisition," *Journal of Speech, Language, and Hearing Research*, vol. 60(11), 3185–3197, 2017
[37] V. Reddy, G. Markova, and S. Wallot, "Anticipatory adjustments to being picked up in infancy," *PLoS ONE,* vol. 8, e65289, 2013.
[38] A. Filipi, "T*oddler and parent interaction: The organisation of gaze, pointing and vocalisation"* (Vol. 192). John Benjamins Publishing, 2009.
[39] E. E. Hilbrink, M. Gattis, and S. C. Levinson, "Early developmental changes in the timing of turn-taking: a longitudinal study of mother-infant interaction," *Frontiers in Psychology*, vol. 6, 1492, 2015.
[40] S. C. Levinson, and F. Torreira, "Timing in turn-taking and its implications for processing models of language," *Frontiers in Psychology*, vol. 6, 731, 2015.
[41] M. Kidwell, "Gaze Shift as an Interactional Resource for Very Young Children, in *Discourse Processes*, vol. 46(2–3), 145–160, 2009.
[42] R. Bakeman, and V. Quera, *Analyzing interaction: Sequential analysis with SDIS and GSEQ*. New York: Cambridge University Press, 1995.
[43] M. Lavelli, C. Barachetti, and E. Florit, "Gesture and speech during shared book reading with preschoolers with specific language impairment," *Journal of Child Language*, vol. 42, 1191–1218, 2015.
[44] C. L. Jr. Webber, and N. Marwan, Ed. *Recurrence Quantification Analysis: Theory and Best Practices*. Cham: Springer, 2015.
[45] N. Marwan, M. C. Romano, M. Thiel, and J. Kurths, "Recurrence plots for the analysis of complex systems," *Physics Reports,* vol. 438(5–6), 237–329, 2007.
[46] C. L. Jr. Webber, and J. P. Zbilut, "Dynamical assessment of physiological systems and states using recurrence plot strategies*," Journal of Applied Physiology,* vol. 76(2), 965–973, 1994.
[47] D. C. Richardson, and R. Dale, "Looking to understand: the coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension," *Cognitive Science,* vol. 29(6), 1045–1060, 2005.
[48] G. Leonardi, I. Nomikou, K. J. Rohlfing, and J. Rączaszek-Leonardi, "Vocal interactions at the dawn of communication: The emergence of mutuality and complementarity in mother-infant interaction," in *Proc. Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob),* Cergy-Pontoise, Paris, 2016, pp. 288–293.
[49] J. Szufnarowska, and K. J. Rohlfing, "Enfolding interaction with two-month-olds," in *Proc. 16th European Conference on Developmental Psychology*. Lausanne, Switzerland, Bologna: Monduzzi Editore, 2014, pp. 213–218.
[50] C. C. Aggrawal, and J. Han, *Frequent Pattern Mining*. Springer, 2014.
[51] E. Suzuki, "Data Mining Methods for Discovering Interesting Exceptions from an Unsupervised Table," *Journal of Universal Computer Science,* vol. 12(6), 627–653, 2006.
[52] G. Dong, and J. Bailey. *Contrast Data Mining: Concepts, Algorithms, and Applications*. Chapman and Hall/CRC, 2012.
[53] N.R. Mabroukeh, and C.I. Ezeife, "A Taxonomy of Sequential Pattern Mining Algorithms," *ACM Computing Surveys,* vol. 43(1), 1–41, 2010.
[54] M. Ruland, "Applying frequent pattern mining to multimodal behaviour in interaction," Unpublished Bachelor thesis. Paderborn University, 2018.
[55] I. Nomikou, K. J. Rohlfing, and J. Szufnarowska, "Educating attention: recruiting, maintaining, and framing eye contact in early natural mother-infant interactions," *Interaction Studies,* vol. 14, 240–267, 2013.
[56] I. Nomikou, M. Koke, and K. J. Rohlfing, "Verbs in mothers' input to 6-month-olds: synchrony between presentation, meaning, and actions is related to later verb acquisition," *Brain Sciences,* vol. 7(5), 52, 2017.
[57] G. Csibra, and G. Gergely, "Natural pedagogy as evolutionary adaptation," *Philosophical Transactions of the Royal Society B: Biological Sciences,* vol. 366(1567), 1149–1157, 2011.
[58] M. Tomasello, and M. J. Farrar, "Joint attention and early language," *Child Development*, vol. 57, 1454–1463, 1986.
[59] K. J. Rohlfing, B. Wrede, A.-L. Vollmer, and P.-Y. Oudeyer, "An alternative to mapping a word onto a concept in language acquisition: pragmatic frames", *Frontiers in Psychology*, vol. 7, 470, 2016.
[60] J. Rączaszek-Leonardi, I. Nomikou, and K. J. Rohlfing, "Young children's dialogical actions: The beginnings of purposeful intersubjectivity," *IEEE Transactions on Autonomous Mental Development*, vol. 3(5), 109–112, 2013.
[61] I. Nomikou, and K. J. Rohlfing, "Language does something. Body action and language in maternal input to three-month-olds," *IEEE Trans. Autonom. Mental Develop.*, vol. 3, no. 2, 113–128, 2011.
[62] L. D'Odorico, R. Cassibba, and N. Salerni, "Temporal relationships between gaze and vocal behavior in prelinguistic and linguistic communication," *Journal of Psycholinguistic Research*, vol. 26(5), 539–556, 1997.

Katharina J. Rohlfing received her Master's in Linguistics, Philosophy, and Media Studies from Paderborn University, Germany, in 1997. As a member of the Graduate Program Task-Oriented Communication, she received her PhD in Linguistics from Bielefeld University in 2002. In 2006, with her interdisciplinary project on the Symbiosis of Language and Action, she became a Dilthey Fellow (Volkswagen Foundation) and Head of the Emergentist Semantics Group at Bielefeld University's CITEC. Currently, she is professor of psycholinguistics at Paderborn University. Her work is on early semantics with a strong interdisciplinary interest in the interface between cognitive development and the early stages of language acquisition.

Giuseppe Leonardi obtained his MA in Psychology from the University of Padua in 1991 and then completed a Doctoral Program in Experimental Psychology at the University of Trieste in 1997. During his doctoral studies, he spent 2 years (1994–1996) at the Center for Complex Systems, Florida Atlantic University, USA. He is now Associate Professor at the University of Economics and Human Sciences in Warsaw (Poland). His work focuses on a dynamical approach to the study of perception–action relations in behavior and interpersonal interactions. This interest also led him to study in depth the methodological challenges this new approach requires and its new analytical applications to empirical data such as recurrence quantification analysis.

Iris Nomikou received her Diploma in Translation Studies from the Ionian University, Corfu, Greece, in 2006, and her MA in Linguistics from Bielefeld University, Germany in 2010. In 2014, she obtained her PhD in Linguistics from Bielefeld University within the project "Symbiosis of Language and Action," supported by the Volkswagen Foundation. She is now a Lecturer at Portsmouth University. Her research interests are in language development and particularly the interactive foundations of language learning. Her research focuses on the interplay between language and bodily experience for the development of meaning.

Joanna Rączaszek-Leonardi received her MA from the University of Warsaw, Warsaw, Poland, and her PhD from the Center of Complex Systems and Brain Sciences at Florida Atlantic University, Boca Raton, Florida, USA. She is a professor at the Faculty of Psychology, University of Warsaw, a cofounder of the interdisciplinary research group Cognitive Systems Warsaw and a head of Human Interaction and Language Lab. Her research interests include basic human interactivity and the nature and role of informational structures, including symbolic ones, such as language, as the means through which this interactivity is controlled.

Eyke Hüllermeier completed his academic education at Paderborn University, Germany, where he received MSc degrees in Business Computing (1993) and Mathematics (1996), a PhD in Computer Science (1997), and his postdoctoral Habilitation (2002). Prior to returning to Paderborn as a full professor in 2014, he spent 2 years as a Marie Curie fellow at the Institut de Recherche en Informatique de Toulouse (France) and held professorships at the Universities of Dortmund, Magdeburg, and Marburg. Currently, he is a full professor in the Department of Computer Science at Paderborn University, where he heads the Intelligent Systems Group. His research interests are centered around theoretical foundations, methods, and applications of artificial intelligence and machine learning. He is Co-Editor-in-Chief of *Fuzzy Sets and Systems* and serves on the editorial board of various other journals including *Machine Learning, Data Mining and Knowledge Discovery* and the *International Journal of Approximate Reasoning*.